Image and Vision Computing 30 (2012) 177–185

Contents lists available at SciVerse ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

The Good, the Bad, and the Ugly Face Challenge Problem[☆]

P. Jonathon Phillips [a,*], J. Ross Beveridge [b], Bruce A. Draper [b], Geof Givens [b], Alice J. O'Toole [c], David Bolme [b], Joseph Dunlop [c], Yui Man Lui [b], Hassan Sahibzada [a], Samuel Weimer [c]

[a] National Institute of Standards and Technology, Gaithersburg, MD 20899, USA
[b] Colorado State University, Fort Collins, CO 46556, USA
[c] The University of Texas at Dallas, Richardson, TX 75083-0688, USA

ARTICLE INFO

Article history:
Received 13 August 2011
Received in revised form 4 January 2012
Accepted 5 January 2012

Keywords:
Face recognition
Challenge problem

ABSTRACT

The Good, the Bad, and the Ugly Face Challenge Problem was created to encourage the development of algorithms that are robust to recognition across changes that occur in still frontal faces. The Good, the Bad, and the Ugly consists of three partitions. The Good partition contains pairs of images that are considered easy to recognize. The base verification rate (VR) is 0.98 at a false accept rate (FAR) of 0.001. The Bad partition contains pairs of images of average difficulty to recognize. For the Bad partition, the VR is 0.80 at a FAR of 0.001. The Ugly partition contains pairs of images considered difficult to recognize, with a VR of 0.15 at a FAR of 0.001. The base performance is from fusing the output of three of the top performers in the FRVT 2006. The design of the Good, the Bad, and the Ugly controls for posevariation, subject aging, and subject "recognizability." Subject recognizability is controlled by having the same number of images of each subject in every partition. This implies that the differences in performance among the partitions are a result of how a face is presented in each image.

Published by Elsevier B.V.

1. Introduction

Face recognition from still frontal images has made great strides over the last twenty years. Over this period, error rates have decreased by three orders of magnitude when recognizing frontal faces in still images taken with consistent controlled illumination in an environment similar to a studio [1–6]. Under these conditions, error rates below 1% at a false accept rate of 1 in 1000 were reported in the Face Recognition Vendor Test (FRVT) 2006 and the Multiple Biometric Evaluation (MBE) 2010 [4,6].

With this success, the focus of research is shifting to recognizing faces taken under less constrained conditions, which include greater variability in pose, ambient lighting, expression, size of the face, and distance from the camera. The trick in designing a face recognition challenge problem is selecting the degree to which the constraints should be relaxed so that the resulting problem has the appropriate difficulty. The complexity of this task is compounded by the fact that it is not well understood how the above factors affect performance. The problem cannot be too easy so that it is merely an exercise in tuning existing algorithms, nor so difficult that progress cannot be made—the three bears problems [2].

Traditionally, a challenge problem is specified by the two sets of images that are to be compared. The difficulty of the problem is then characterized by the performance of a set of algorithms tasked with matching the two sets of face images. To create a problem with the desired level of difficulty, the performance of a set of algorithms can be one component in the design process. Others factors in the selection process include limiting the number of images per person and requiring pairs of images of a person to be collected on different days.

The Good, the Bad, and the Ugly (GBU) challenge problem consists of three partitions called the Good, the Bad, and the Ugly. The difficulty of each partition is based on the performance of three top performers in the FRVT 2006. The Good partition consists of pairs of face images of the same person that are easy to match; the Bad partition contains pairs of face images of a person that have average matching difficulty; and the Ugly partition concentrates on difficult to match face pairs. Nominal performance on the GBU is based on fusing the results from three top performers in the FRVT 2006. The Good partition has a verification rate (VR) of 0.98 at a false accept rate (FAR) of 0.001. The Bad and Ugly partitions have VRs of 0.80 and 0.15 at FAR of 0.001, respectively. The performance range over the three partitions is roughly an order of magnitude.[1]

[☆] This paper has been recommended for acceptance by special issue Guest Editors Rainer Stiefelhagen, Marian Stewart Bartlett and Kevin Bowyer.

[*] Corresponding author. Tel.: +1 301 975 5348; fax: +1 301 9755348.
E-mail address: jonathon@nist.gov (P.J. Phillips).

[1] Instructions for obtaining the complete GBU distribution can be found at http://face.nist.gov. Instructions for obtaining the LRPCA algorithm can be found at http://www.cs.colostate.edu/facerec.

0262-8856/$ – see front matter. Published by Elsevier B.V.
doi:10.1016/j.imavis.2012.01.004

There are numerous sources of variation, known and unknown, in face images that can affect performance. Four of these factors were explicitly controlled for in the design of the GBU challenge problem: subject aging, pose, change in camera, and variations among faces. The data collection protocol eliminated or significantly reduced the impact of three of these factors. Changes in the appearance of a face due to aging is not a factor because all images were collected in the same academic year (9 month time span). However, the data set contains the natural variations in a person's appearance that would occur over an academic year. Because all the images were collected by the same model of camera, difference in performance cannot be attributable to changes in the camera. Changes in pose are not a factor because the data set consists of frontal face images.

One potential source of variability in performance is that people vary in their "recognizability." To control for this source of variability, there are face images of each person in all three partitions. In addition, each partition has the same number of images of each person. Because the partition design controls for variation in the recognizability of faces, differences in performance among the three partitions are a result of how the faces are presented.

The primary goal of the GBU face challenge problem is to focus attention on the fundamental problem of comparing single frontal face images across changes in appearance. There are numerous applications that require matching single frontal images. Examples include recognition from face images on mug shots, passports, driver's licenses, and US Government PIV Identity cards. Additionally, there are numerous commercial and academic organizations that have developing algorithms specifically for this application area. The structure of the GBU encourages algorithm development in these application areas as well as supports both the development of new recognition algorithms and experiments to identify factors that affect performance.

2. Generation of the Good, the Bad, and the Ugly partitions

The GBU partitions were constructed from the Notre Dame multi-biometric data set used in the FRVT 2006 [4]. The images for the partitions were selected from a superset of 9307 images of 570 subjects. All the images in the superset are frontal still face images collected either outside or with ambient lighting in hallways. The images were acquired with a 6 Mega-pixel Nikon D70 camera. All photos were taken in the 2004–2005 academic year (Aug 2004 through May 2005).

Each partition in the GBU is specified by two sets of images: a target set and a query set. For each partition, an algorithm computes a similarity score between all pairs of images in that partition's target and query sets. A similarity score is a measure of the similarity between two faces. Higher similarity scores imply greater likelihood that the face images are of the same person. If an algorithm reports a distance measure, then a smaller distance measure implies greater likelihood that the face images are of the same person. Distances are converted to similarity scores by multiplying by negative one. The set of all similarity scores between a target and a query set is called a similarity matrix. A pair of face images of the same person is called a match pair, and a pair of face images of different people is called a non-match pair. From the similarity matrix, receiver operating characteristics (ROC) and other measures of performance can be computed.

To construct the GBU Challenge Problem we sought to specify target and query sets for each of the three partitions such that recognition difficulty would vary markedly while at the same time factors such as the individual people involved or number of images per person remained the same. To gauge the relative difficulty associated with recognizing a pair of images, similarity scores were created by fusing scores from three of the top performing algorithms in the FRVT 2006; this fusion process is described more fully in the next section.

The following constraints were imposed when selecting the GBU partitions:

Distinct Images: An image can only be in one target or query set.

Balanced subject counts: The number of images per person is the same in all target and query sets.

Different days: The images in all match pairs were taken on different days.

After applying these constraints, and given the total number of images available, the number of images per person in the target and query sets was selected to fall between 1 and 4. This number depended upon the total availability of images for each person.

The selection criteria for the partition results in the following properties. An image is only in one partition. There are the same number of match face pairs for each subject in each partition. There are the same number of non-match pairs between any two subjects in each partition. This implies that any difference in performance between the partitions is not a result of different people. The difference in performance is a result of the different conditions under which the images were acquired. Figs. 1, 2, and 3, show examples of matching face pairs (mated vertically) from each of the partitions.

The images included in the GBU target and query sets were decided independently for each person. For each subject *i*, a subject-specific similarity matrix S_i is extracted from a larger matrix containing similarity scores from the FRVT2006 fusion algorithm. Each subject-specific matrix contained all similarity scores between pairs of images of subject *i*. For the Good partition, a greedy selection algorithm iteratively added match face pairs for subject *i* that maximized the average similarity score for subject *i*; for the Ugly partition, match face pairs were selected to minimize the average similarity score for subject *i*; and for the Bad partition, face pairs for subject *i* were selected to maintain an approximately average similarity score. The selection process for each subject was repeated until the desired number of images were selected for that subject. Since the images for each subject are selected independently, the similarity score associated with a good face pair can vary from subject to subject (similarly for the Bad and Ugly partitions).

Each of the GBU target and query sets contains 1085 images for 437 distinct people. The distribution of image counts per person in the target and query sets are 117 subjects with 1 image, 122 subjects with 2 images, 68 subjects with 3 images, and 130 subjects with 4 images. In each partition there is 3297 match face pairs and 1,173,928 non-match face pairs. In the GBU image set 58% of the subjects are male and 42% female; 69% of the subjects are Caucasian, 22% east Asian, 4% Hispanic, and the remaining 5% other groups; and 94% of the subjects are between 18 and 30 years old with the remaining 6% over 30 years old. For the images in the GBU, the average distance between the centers of the eyes is 175 pixels with a standard deviation of 36 pixels.

3. The FRVT 2006 fusion performance

Performance results for the GBU Challenge Problem are reported for the GBU FRVT 2006 fusion algorithm, which is a fusion of three of the top performers in the FRVT 2006. The algorithms were fused in a two-step process. First, for each algorithm the median and the median absolute deviation (MAD) were estimated from every 1 in 1023 similarity scores (median_k and MAD_k are the median and MAD for algorithm *k*). The median and MAD were estimated from 1 in 1023 similarity scores to avoid over tuning the estimates to the data. The similarity scores were selected to evenly sample the images in the experiment.[2] Second, the fused similarity scores are the sum of the individual algorithm similarity scores after the median has been subtracted and then divided by

[2] The parameters for the fusion formula were computed from a subset of the similarity scores rather than on the complete set of similarity scores. This was done with the goal of generating a fusion formula that would generalize to additional faces or algorithm data, rather than being overly tuned to this particular dataset. In the algorithm evaluations carried out by NIST, the commonly applied procedure is to combine data with a method that has the ability to generalize.

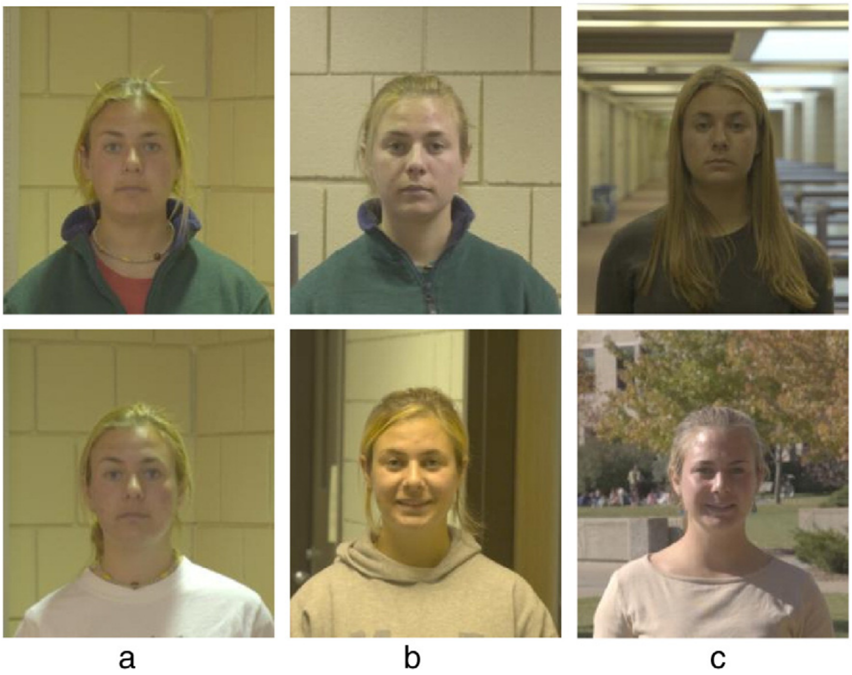

Fig. 1. Examples of face pairs of the same person from each of the partitions: (a) good, (b) challenging, and (c) very challenging.

the MAD. If s_k is a similarity score for algorithm k and s_f is a fusion similarity score, then $s_f = \sum_k (s_k - \text{median}_k)/\text{MAD}_k$.

Fig. 4 reports performance of the fusion algorithm on each of the partitions. Fig. 5 shows the distribution of the match and non-matches for the fusion algorithm on all three partitions. The non-match distribution is stable across all three partitions. The match distribution shifts for each partition. The Good partition shows the greatest difference between the median of the match and non-match distributions and the Ugly partition shows least difference.

4. Protocol

The protocol for the GBU Challenge Problem is one-to-one matching with training, model selection, and tuning completed prior to computing performance on the partitions. Consequently, under this protocol, the similarity score $s(t, q)$ between a target image t and a query image q does not in any way depend on the other images in the target and query sets. Avoiding hidden interactions between images other than the two being compared at the moment provides the clearest picture

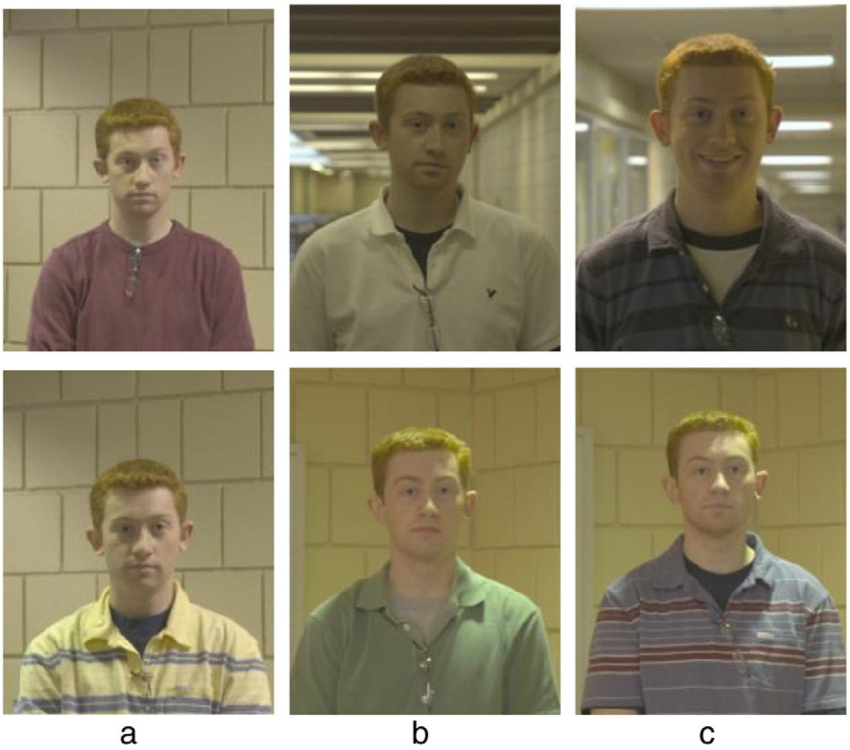

Fig. 2. Examples of face pairs of the same person from each of the partitions: (a) good, (b) challenging, and (c) very challenging.

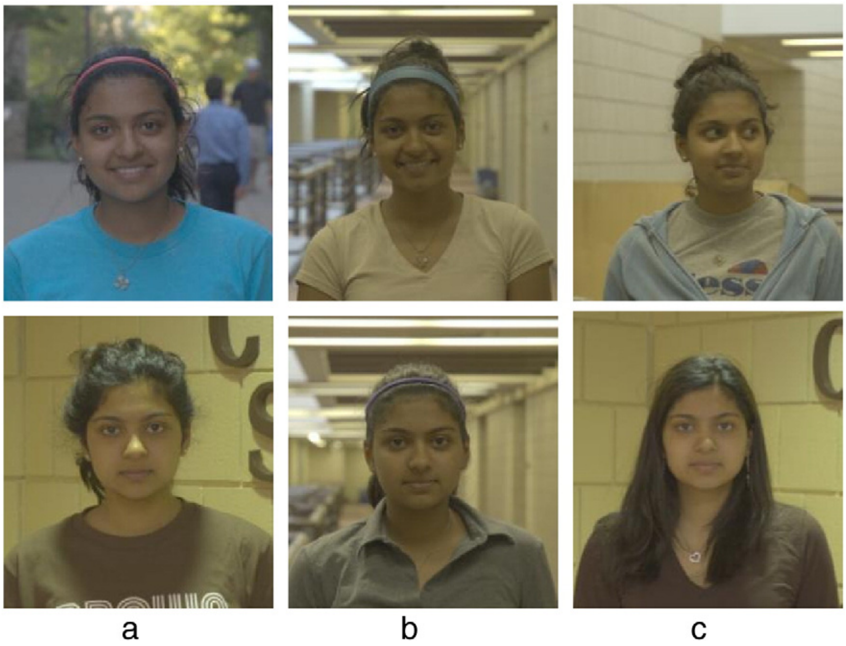

Fig. 3. Examples of face pairs of the same person from each of the partitions: (a) good, (b) challenging, and (c) very challenging.

of how algorithms perform. More formally, any approach that redefines similarity $s(t, q; \mathcal{T})$ such that it depends upon the target (or query) image set $\mathcal{T}$ is NOT allowed in the GBU Challenge Problem.

To maintain separation of training and test sets, an algorithm cannot be trained on images of any of the subjects in the GBU Challenge Problem. It is important to note that there are images of the subjects in the GBU problem that are in the FRGC and the MBGC data sets. These images must be excluded from model selection, training, or tuning of an algorithm.

We illustrate acceptable and unacceptable training protocols with three examples. The first example is the training of a principal components analysis (PCA) based face-recognition algorithm. In the algorithm, PCA is performed on a training set to produce a set of Eigenfaces. A face is represented by projecting a face image on the set of Eigenfaces. To meet the training requirements of the protocol, images of subjects in the GBU must be excluded from the PCA decomposition that produces a set of Eigenfaces. The benchmark algorithm in Section 5 includes a training set that satisfies the training protocol.

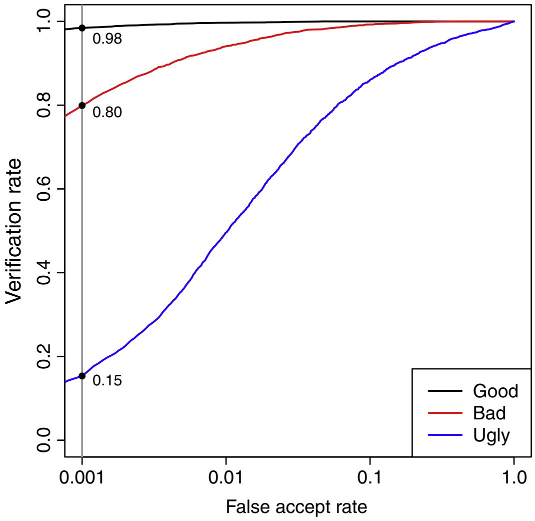

Fig. 4. ROC for the Fusion algorithm on the Good, the Bad, and the Ugly partitions. The verification rate for each partition at a FAR of 0.001 is highlighted by the vertical line at FAR = 0.001.

A second example is taken from a common training procedure for linear discriminant analysis (LDA) in which the algorithm is trained on the images in a target set. Generally, it is well known that the performance of algorithms can improve with such training, but the resulting levels of performance typically do not generalize. For example, we've conducted experiments with an LDA algorithm trained on the GBU target images and performance improved over the baseline algorithm presented, see Section fsec:Benchmark Algorithm. However, when we trained our LDA algorithm following the GBU protocol, performance did not match the LDA algorithm trained on a GBU target set.

The GBU protocol does permit image specific representations as long as the representation does not depend on other images of other subjects in the GBU Challenge Problem. An example is an algorithm based on person-specific PCA representations. In this example, during the geometric normalization process, 20 slightly different normalized versions of the original face would be created. A person-specific PCA representation is generated from the set of 20 normalized face images. This method conforms with the GBU training protocol because the 20 face images and the person specific PCA representation are functions of the original single face image. When there are multiple images of a person in a target or query set, this approach will generate multiple image-specific representations. This training procedure does not introduce any dependence upon other images in the target set and consequently is permitted by the GBU protocol.

5. Baseline algorithm

The GBU Challenge Problem includes a baseline face recognition algorithm as an entry point for researchers. The baseline serves two purposes. First, it provides a working example of how to carry out the GBU experiments following the protocol. This includes training, testing and evaluation using ROC analysis. Second, it provides a performance standard for algorithms applied to the GBU Challenge Problem.

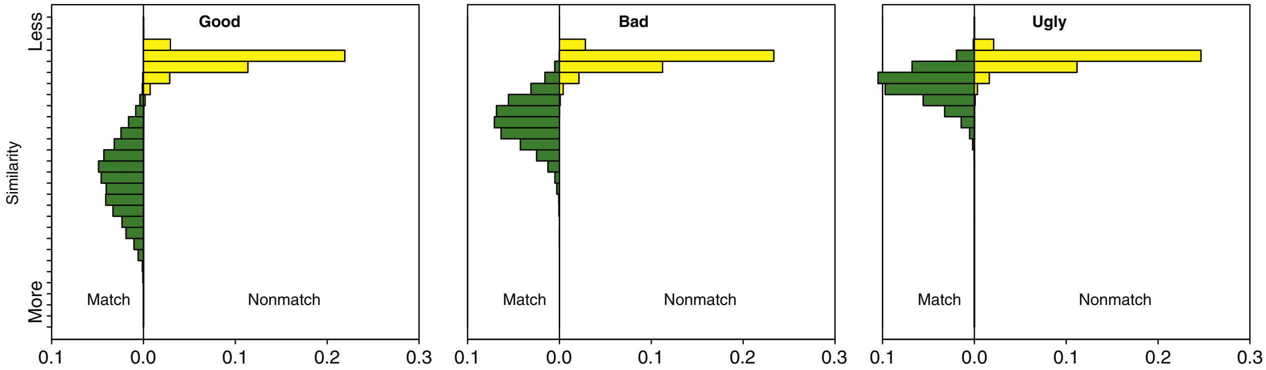

Fig. 5. Histogram of the match and non-match distributions for the Good, the Bad, and the Ugly partitions. The green bars represent the match distribution and the yellow bars represent the non-match distribution. The horizontal axes indicate relative frequency of similarity scores.

The architecture of the baseline algorithm is a refined implementation of the standard PCA-based face recognition algorithm, also known as Eigenfaces [7,8]. These refinements considerably improve performance over a standard PCA-based implementation. The refinements include representing a face by local regions, a self quotient normalization step, and weighting eigenfeatures based on Fischer's criterion. We refer to the GBU baseline algorithm as local region PCA (LRPCA).

It may come as a surprise to many in the face recognition community that a PCA-based algorithm was selected for the GBU benchmark algorithm. However, when developing the LRPCA baseline algorithm, we explored numerous standard alternatives, including LDA-based algorithms and algorithms combining Gabor based features with kernel methods and support vector machines. For performance across the full range of the GBU Challenge Problem, our experiments with alternative architectures have not resulted in overall performance better than the LRPCA baseline algorithm.[3]

5.1. A step-by-step algorithm description

The algorithm's first step is to extract a cropped and geometrically-normalized face region from an original face image. The original image was assumed to be a still image and the pose of the face is close to frontal. The face region in the original is scaled, rotated, and cropped to a specified size and the centers of the eyes are horizontally aligned and placed on standard pixel locations. Scaling, rotating, and cropping of the face is based on the centers of the eyes which were manually located.[4] In the baseline algorithm, the face chip is 128 by 128 pixels with the centers of the eyes spaced 64 pixels apart. The baseline algorithm runs in two modes: partially and fully automatic. In the partially automatic mode the coordinates of the centers of the eyes are provided; in the fully automatic mode, the centers of the eyes are located by the baseline algorithm.

In the LRPCA algorithm, the PCA representation is based on 14 local regions. The 14 regions include the complete face chip. The 14 local regions are cropped out of a normalized face image. Some of the local regions overlap, see Fig. 6. The local regions are centered relative to the average location of the eyes, eyebrows, nose and mouth.

The next step normalizes the 14 face regions to attenuate variation in illumination. First, self quotient normalization is independently applied to each of the 14 regions [9]. The self quotient normalization procedure first smoothes each region by convolving it with a two-dimensional Gaussian kernel and then divides the original region by the smoothed region, see Fig. 7. In the final normalization step, the pixel values in each region are further adjusted to have a sample mean of zero and a sample standard deviation of one.

During training, 14 distinct PCA subspaces are constructed, one for each of the face regions. From each PCA decomposition, the 3rd through 252th eigenvectors are retained to represent the face. The decision to use these eigenvectors was based upon experiments on images similar to the images in the GBU Challenge Problem. A region in a face is encoded by the 250 coefficients computed by projecting the region onto the region's 250 eigenvectors. A face is encoded by concatenating the 250 coefficients for each of the 14 regions into a new vector of length 3500.

Each dimension in the PCA subspace is further scaled. First, the representation is whitened by scaling each dimension to have a sample standard deviation of one on the training set. Next, the weight on each dimension is further adjusted based on Fisher's criterion, which is the ratio of the between class variance and the within class variance (σ_b^2/σ_w^2). This weight is computed based on the images in the training set emphasizes the dimensions along which images of different people are spread apart and attenuates the dimensions along which the average distance between images of the same person and images of different people are roughly the same.

During the recognition process, images are first processed as described above and then projected into the 14 distinct PCA subspaces associated with each of the 14 regions. The coordinates of images

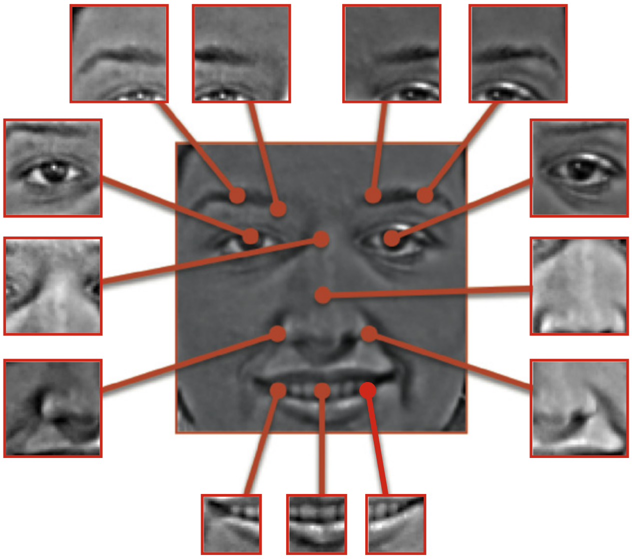

Fig. 6. This figure shows the 14 local regions in the LRPCA algorithm. The fourteen regions include the cropped face. The crop face has been geometrically normalized and the self quotient procedure performed.

[3] This statement was accurate for the original submission to the IEEE Ninth International Conference on Automatic Face and Gesture Recognition 2011.

[4] The coordinates of the manually located centers of the eyes are made available to researchers.

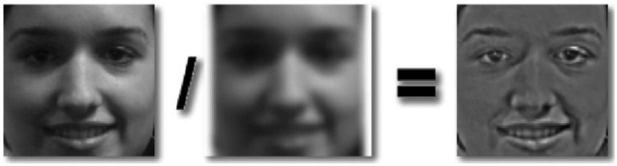

Fig. 7. This figure illustrates the computation of a self-quotient face image. The face image to the left is a cropped and geometrically normalized image. The image in the middle is the geometrically normalized image blurred by a Gaussian kernel. The image on the left is a self-quotient image. This image is obtained by pixel-wise division of the normalized image by the blurred image.

projected into these spaces, 250 for each of the 14 regions, are then concatenated into a single feature vector representing the appearance of that face. This produces one vector per face image; each vector contains 3500 values. The baseline algorithm measures similarity between pairs of faces by computing the Pearson's correlation coefficient between pairs of these vectors. The performance of the baseline algorithm on the GBU Challenge Problem is summarized in Fig. 8. A comparison of performance of the Fusion and the LRPCA-baseline algorithm is given in Table 1.

A recent area of interest in face recognition and biometrics is recognition from the ocular region of the face. There is interest in recognition from both near infrared and visible imagery. The region-based design of the LRPCA algorithm allows for baselining ocular performance on the GBU partitions. Baseline performance for the left ocular is computed from three of the 14 regions. The regions are the left eye and two left eye brow regions. For the right ocular region, performance is computed from the right eye and two right eye brow regions. Here left and right are defined in respect to the subject where the left ocular region corresponds to a subject left eye. Performance for the LRPCA-ocular baseline for the left and right ocular regions is given in Fig. 9.

A summary of performance of the Fusion, the LRPCA-face baseline and the LRPCA-ocular baseline algorithms are given in Table 1.

6. Analysis

The goals of the GBU includes understanding the properties of face recognition algorithms. One aspect is understanding and characterizing the factors that affect performance.

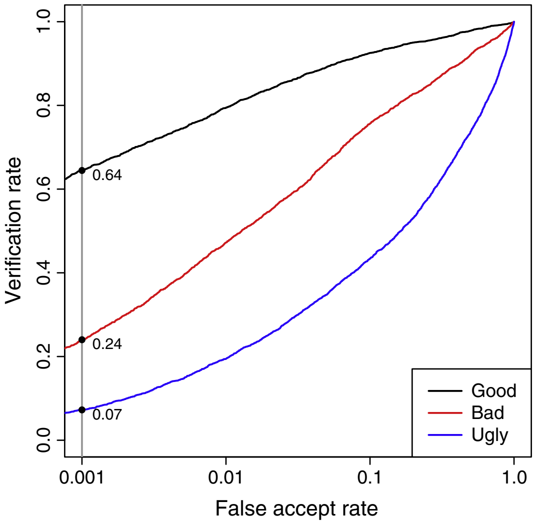

Fig. 8. ROC for the LRPCA baseline algorithm on the GBU partitions. The verification rate for each partition at a FAR of 0.001 is highlighted by the vertical line at FAR = 0.001.

Table 1
Performance of the Fusion, the LRPCA-face baseline, and the LRPCA-ocular baseline algorithms. For the ocular baseline, performance is given for both the left and the right ocular regions. The verification rates at a FAR = 0.001 are given.

Partition	Fusion	LRPCA-face	LRPCA-ocular	
			Left	Right
Good	0.98	0.64	0.47	0.46
Bad	0.80	0.24	0.16	0.17
Ugly	0.15	0.07	0.05	0.05

6.1. Limits of expression and lighting effects

A common assumption in the face recognition community is that the primary factors affecting performance are aging, pose, illumination and expression (A-PIE). Here we show that this assumption does not apply to the GBU. Aging, which refers to the elapsed time between two face images of a person, is not a factor because all the images in the GBU were collected with a nine month time span. Additionally, all images in the GBU nominally have a frontal pose and therefore it is assumed pose is not a factor. There are a few images in the GBU that are clearly not frontal, but the number is sufficiently small that it does not account for the difference in performance among the three partitions.

Of the four A-PIE factors, the GBU contains significant variations in illumination and expression. Both these factors can have a significant impact on performance.

In a meta-analysis on the effect of covariates on face recognition performance, changes in expressions were identified as a factor that consistently impacted performance [10]. For expression, the data collection protocol requested that subjects present neutral and smiling expressions. Face image pairs are categorized as having the same expression when the two images are labeled smiling or two images are labeled neutral. Similarly, pairs labeled as different expressions when one image is labeled smiling and the other is labeled neutral.

Lighting is one factor affecting face recognition performance and has been extensively studied [11]. According to the spherical harmonic theory [12], 99% of the reflected energy can be captured by the first nine spherical harmonics. An image can therefore be relighted based on a set of lighting images. Sim and Kanade [13] applied a Bayesian model and a kernel regression technique to synthesize new images for different illuminations. This work was extended to estimate the illumination directions in facial imagery [14]. The output of this analysis was the dominant lighting direction. For this study, lighting direction was estimate by the method of Beveridge et al. [14]. The lighting direction was quantized to frontal, right, left or from behind. For the analysis in this section, a pair of face images has the same lighting if both images had the same lighting direction; otherwise, the face pair had different lighting.

To assess the impact of change in expression and lighting direction, all match face pairs are given one of labels: 1) same lighting, same expression; 2) same lighting, different expression; 3) different lighting, same expression; and 4) different lighting, different expression. Fig. 10 summarizes the results of this analysis. In Fig. 10 there is a bar for each partition, and each bar consists of four color coded rectangles. Each of the rectangles corresponds to one of the four labels regarding the status of change in expression and lighting direction. The length of the rectangles is proportional to the number match face pairs with the corresponding label. For the Good partition, there are 1641 face pairs with the same lighting and same expression, 343 face pairs with different lighting and different expression. Also reported for each change in expression and lighting condition, is the verification rate at aFAR of 1/1000. The easiest case is the same lighting and same expression for the Good partition with a verification rate of 99.5%. The most difficult case is different lighting and different expression for the Ugly partition with a verification rate of 7.6%.

The Good partition contained the most number of same lighting, same expression match face pairs, followed by the Bad and then the

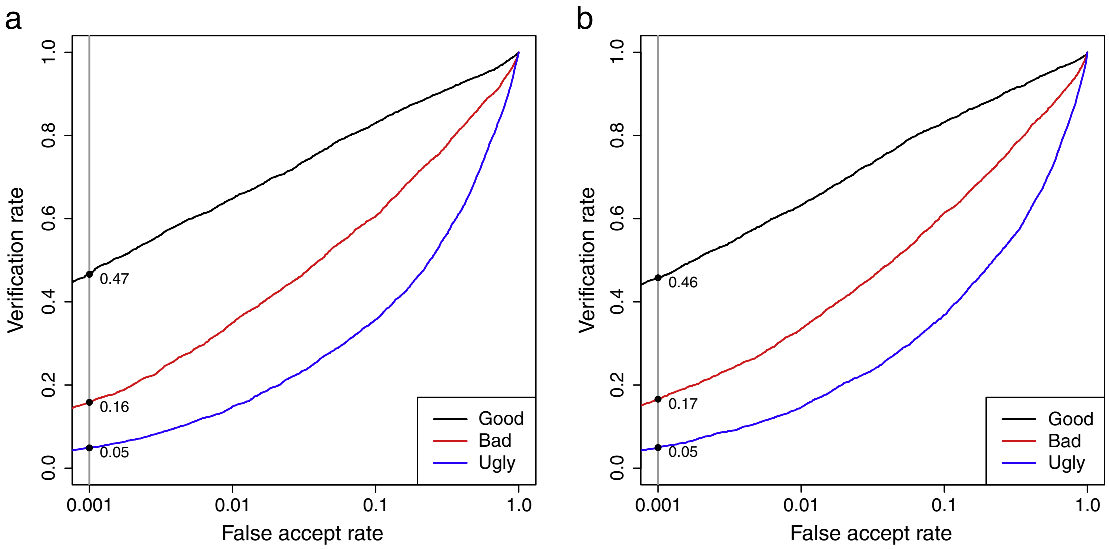

Fig. 9. ROC for the LRPCA-ocular baseline algorithm on the Good, the Bad, and the Ugly partitions. In (a) performance is for the left ocular region that consists of the left eye and two left eye-brow regions; performance in (b) is for corresponding right ocular regions. The verification rate for each partition at a FAR of 0.001 is highlighted by the vertical lines at FAR = 0.001.

Ugly partitions. Similarly, the Ugly partition contained the most number of different lighting, different expression match pairs, followed by the Bad and then the Good partitions.

For the Good partition, the verification rate at a FAR of 1/1000 for all four cases is better than all cases in the Bad partition. Similarly, performance on all cases for the Bad partition is better than all cases on the Ugly partition. If changes in lighting or expression were the source of a large percentage of the difference in performance among the three partitions, then there would not be the observed stratification of performance.

The results in this section suggest that there exist other factors that significantly affect performance. Currently these factors have not been identified and the GBU challenge problem provides a basis for identifying these factors.

6.2. Zoo analysis

One of the open questions in face recognition is "are some faces harder or easier to recognize than other faces?" This question implicitly implies that faces are ordered from hardest to easiest to recognize. If there is an ordering to the recognizability of faces, then the order should be stable over imaging conditions. Because there are the same number of images of each person in each the GBU partitions, the GBU is amenable to examining the relative difficulty of faces across imaging conditions.

Characterizing the ease or difficulty of recognizing a person from their face, or more generally from any biometric, is referred to as the *biometric zoo problem*. Zoo analysis falls into two basic classes. The first looks at the "most" and "least" difficult to recognize faces–the extremes of the distribution [15–17]. The second looks at the relative ranks of all the faces in a data set [18,19]. We chose to follow the second method.

Our analysis is based on two measures of face difficulty. The first is the median match score of a subject. The second is the median non-match scores of a subject. The degree of consistency in recognizability of the faces in the GBU among the partitions is measured by Spearman's correlation coefficient.

For a partition, the median match score for subject k is the median of all the match similarity scores where both the target and query are images of subject k. Likewise, the median non-match score for subject k is median of all non-match similarity scores where either the target or query image is an image of subject k.

Fig. 11 plots the correlation between the median subject match and non-match scores among the three partitions. Table 2 reports Spearman's correlation coefficient corresponding to the scatterplots in Fig. 11.

For the median match scores, correlation varies from 0.32 to 0.46. This shows some degree of similarity in the relative ranking among the partitions for the median match scores. For the median non-match scores, correlation varies from 0.03 to 0.38. The 0.03 coefficient

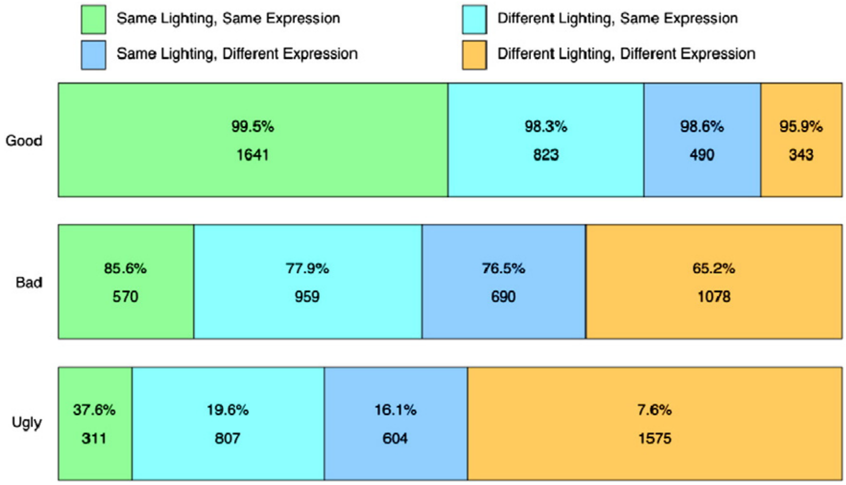

Fig. 10. Effect of expression and lighting over the GBU partitions. Each bin shows the verification rate at 1/1000 FAR and the total number of match pairs.

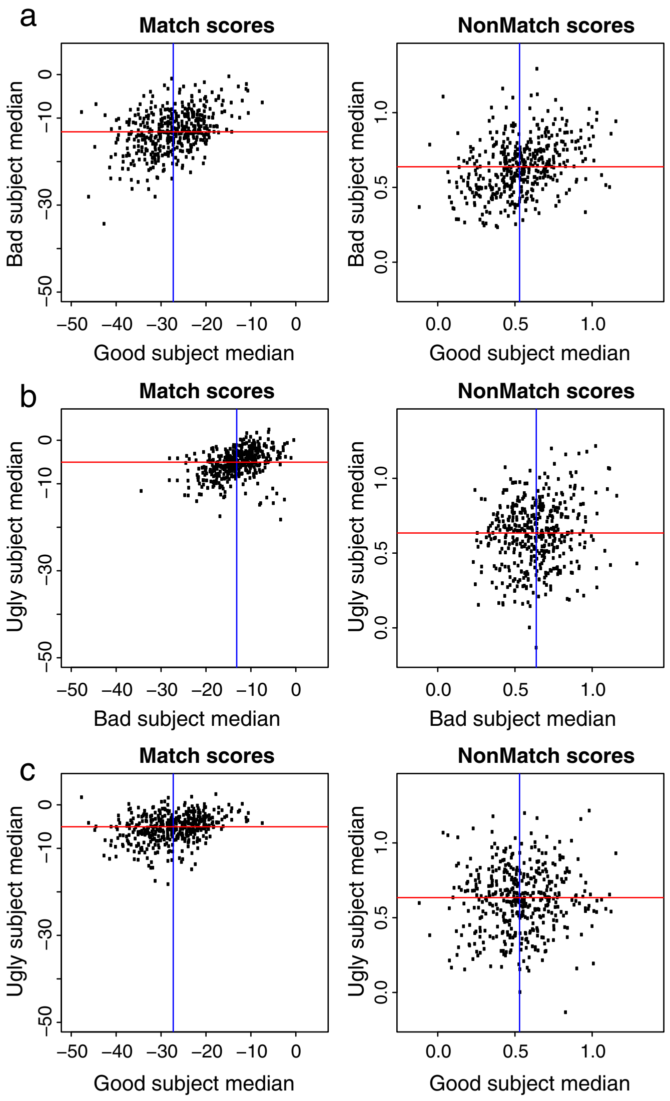

Fig. 11. Scatterplots of median subject match scores and median subject non-match scores. Each dot corresponds to a subject. The blue vertical line marks the median subject match (non-match) scores for the horizontal axis. The red horizontal line marks the median subject match (non-match) scores for the vertical axis. (a) Compares median subject match and non-match between the Good and Bad partitions; (b) compares median subject match and non-match between the Bad and Ugly partitions; and (c) compares median subject match and non-match between the Good and Ugly partitions.

is between the Good and Ugly partitions, which is essentially random. The median non-match scores affects the ability of a person to be consistently impersonated. This result suggests that the ease with which a person can be impersonated varies with image acquisition conditions. In addition, the result suggests that for non-matches a zoo structure does not exist across changes in imaging conditions.

7. Discussion and conclusion

This paper introduces the Good, the Bad, and the Ugly Challenge Problem. The main goal of the challenge is to encourage the development of robust algorithms for recognizing frontal faces taken outside of studio style image collections. The three partitions in the GBU Challenge Problem emphasize the range of performance that is possible when comparing faces photographed under these conditions. This structure allows for researchers to concentrate on the “hard” aspects of the problem while not compromising performance on the “easier” aspects.

Table 2
Spearman's correlation coefficients for zoo analysis.

	Median match scores	Median non-match scores
Good–Bad	0.39	0.38
Bad–Ugly	0.46	0.11
Good–Ugly	0.32	0.03

Partitioning the challenge by levels of difficulty is the most prominent feature of the GBU Challenge Problem design. Another is controlling for the “recognizability” of people by selecting images of the same 437 people for inclusion in each of the GBU partitions. The data in the three partitions is further balanced so as to ensure that for each person the number of target and query images in each partition is the same. The design of the GBU Challenge Problem means that any difference in performance observed between partitions cannot be attributed to differences between people or numbers of images for individual people.

The unique design of the GBU Challenge Problem allows researchers to investigate factors that influence the performance of algorithms. O'Toole et al. [20,21] looked at the demographic effects on the nonmatch distribution. Beveridge et al. [22] showed that the quality of face images comes in pairs. Quality comes in pairs was shown by the existence of contrary face image: images that have a contrary nature because they simultaneously have high and low quality. Additional possible lines of investigation include understanding the factors that characterize the difference in match face pairs across the partitions. Our results show that changes in expression and lighting direction do not characterize the majority of the variation across the partitions. A second line of research is characterizing the recognizability of a face; e.g., the biometric zoo. Our zoo experiments suggest that recognizability is not stable across the partitions. A third line of research is developing methods for predicting performance of face recognition algorithms. The design of the GBU Challenge Problem encourages both the development of algorithms and the investigation of methods for understanding algorithm performance.

Acknowledgments

The authors wish to thank the Federal Bureau of Investigation (FBI) and the Technical Support Working Group (TSWG) for their support of this work. The identification of any commercial product or trade name does not imply endorsement or recommendation by the Colorado State University, the National Institute of Standards and Technology or the University of Texas at Dallas. The authors thank Jay Scallan for his assistance in preparing the GBU challenge problem.

References

[1] P.J. Phillips, H. Wechsler, J. Huang, P. Rauss, The FERET database and evaluation procedure for face-recognition algorithms, Image Vision Comput. J. 16 (1998) 295–306.
[2] P.J. Phillips, H. Moon, S. Rizvi, P. Rauss, The FERET evaluation methodology for face-recognition algorithms, IEEE Trans. Pattern Anal. Mach. Intell. 22 (2000) 1090–1104.
[3] P.J. Phillips, P.J. Flynn, T. Scruggs, K.W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, W. Worek, Overview of the face recognition grand challenge, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005, pp. 947–954.
[4] P.J. Phillips, W.T. Scruggs, A.J. O'Toole, P.J. Flynn, K.W. Bowyer, C.L. Schott, M. Sharpe, FRVT 2006 and ICE 2006 large-scale results, IEEE Trans. Pattern Anal. Mach. Intell. 32 (2010) 831–846.
[5] P.J. Phillips, P.J. Flynn, J.R. Beveridge, W.T. Scruggs, A.J. O'Toole, D. Bolme, K.W. Bowyer, B.A. Draper, G.H. Givens, Y.M. Lui, H. Sahibzada, J.A. Scallan III, S. Weimer, Overview of the Multiple Biometrics Grand Challenge, Proceedings Third IAPR International Conference on Biometrics, 2009.
[6] P.J. Grother, G.W. Quinn, P.J. Phillips, MBE 2010: Report on the evaluation of 2D still-image face recognition algorithms, NISTIR 7709, National Institute of Standards and Technology, 2010.
[7] M. Turk, A. Pentland, Eigenfaces for recognition, J. Cogn. Neurosci. 3 (1991) 71–86.
[8] M. Kirby, L. Sirovich, Application of the Karhunen–Loeve procedure for the characterization of human faces, IEEE Trans. Pattern Anal. Mach. Intell. 12 (1990) 103–108.
[9] H. Wang, S. Li, Y. Wang, J. Zhang, Self quotient image for face recognition, Proceedings, International Conference on Image Processing, volume 2, 2004, pp. 1397–1400.

[10] Y.M. Lui, D. Bolme, B.A. Draper, J.R. Beveridge, G. Givens, P.J. Phillips, A meta-analysis of face recognition covariates, IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems, Washington, DC, 2009.

[11] X. Zou, J. Kittler, K. Messer, Illumination invariant face recognition: A survey, IEEE International Conference on Biometrics: Theory, Applications, and Systems, 2007.

[12] R. Basri, D. Jacobs, Lambertian reflectance and linear subspaces, IEEE Trans. Pattern Anal. Mach. Intell. 25 (2003) 218–233.

[13] T. Sim, T. Kanade, Combining models and exemplars for face recognition: an illuminating example, CVPR Workshop on Models versus Exemplars in Computer Vision, 2001.

[14] J.R. Beveridge, D.S. Bolme, B.A. Draper, G.H. Givens, Y.M. Lui, P.J. Phillips, Quantifying how lighting and focus affect face recognition performance, Proceedings IEEE Computer Society and IEEE Biometrics Council Workshop on Biometrics, 2010.

[15] G. Doddington, W. Ligget, A. Martin, M. Przybocki, D. Reynolds, Sheeps, goats, lambs, and wolves: a statistical analysis of speaker performance in the NIST 1998 recognition evaluation, Proceedings ICSLP '98, 1998.

[16] N. Yager, T. Dunstone, The biometric menagerie, IEEE Trans. Pattern Anal. Mach. Intell. 32 (2010) 220–230.

[17] M.N. Teli, J.R. Beveridge, P.J. Phillips, G.H. Givens, D.S. Bolme, B.A. Draper, Biometric zoos: theory and experimental evidence, International Joint Conference on Biometrics, 2011.

[18] N. Poh, S. Bengio, A. Ross, Revisiting Doddington's zoo: a systematic method to assess user-dependent variabilities, Proc. of Second Workshop on Multimodal User Authentication (MMUA), 2006.

[19] A. Ross, A. Rattani, M. Tistarelli, Exploiting the “Doddington zoo” effect in biometric fusion, IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems, 2009.

[20] A.J. O'Toole, P.J. Phillips, X. An, J. Dunlop, Demographic effects on estimates of automatic face recognition performance, Proceedings, Ninth International Conference on Automatic Face and Gesture Recognition, 2011.

[21] A.J. O'Toole, P.J. Phillips, X. An, J. Dunlop, Demographic effects on estimates of automatic face recognition performance, Image and Vision Computing Journal (this issue).

[22] J.R. Beveridge, P.J. Phillips, G.H. Givens, B.A. Draper, M.N. Teli, D.S. Bolme, When high-quality face images match poorly, Proceedings, Ninth International Conference on Automatic Face and Gesture Recognition, 2011.